

# Determining the Privacy-loss Budget

## Research into Alternatives to Differential Privacy

National Advisory Committee  
May 27, 2021

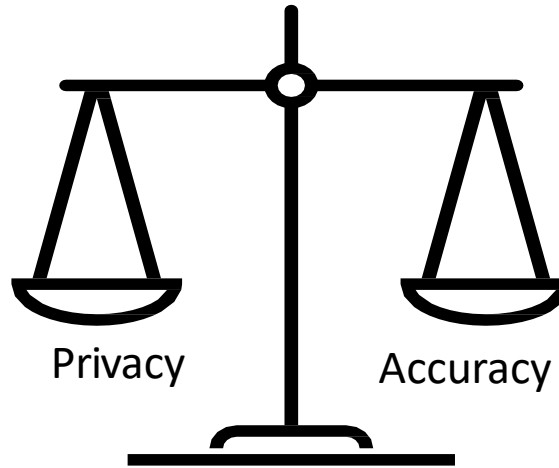
Michael Hawes  
Rolando A. Rodríguez  
U.S. Census Bureau

Pre-decisional

Shape  
your future  
START HERE >

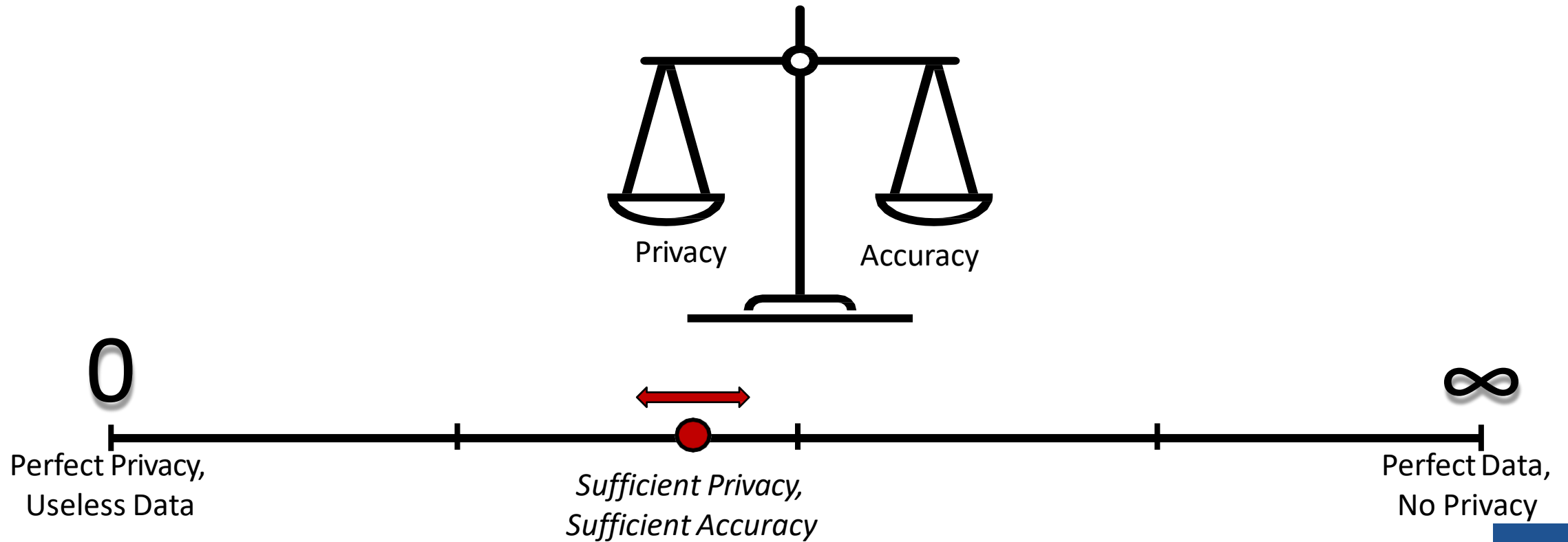
United States<sup>®</sup>  
**Census**  
**2020**

# What is a Privacy-loss Budget?



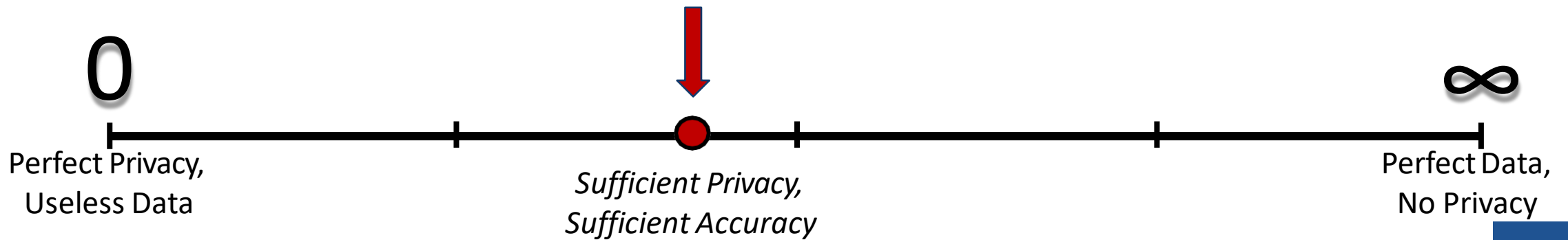
Any disclosure avoidance mechanism imposes a fundamental tradeoff between data protection (privacy/confidentiality) and data accuracy/fitness-for-use.

# What is a Privacy-loss Budget?



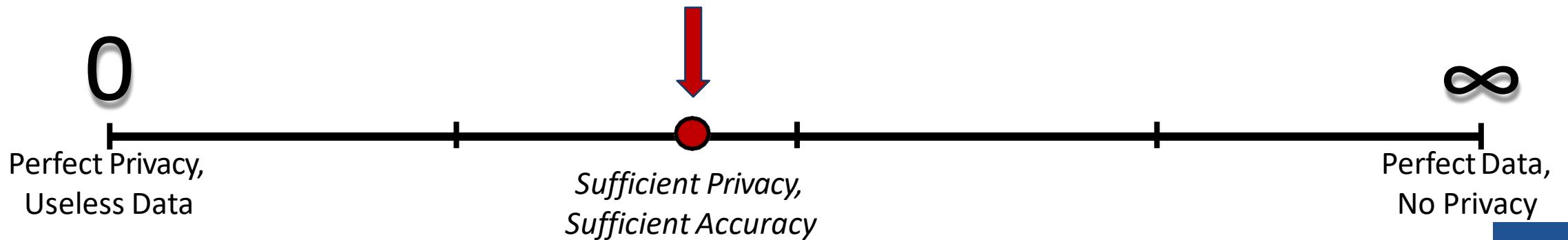
# What is a Privacy-loss Budget?

**Privacy-loss Budget**  
(PLB, " $\epsilon$ ", " $\rho$ ")



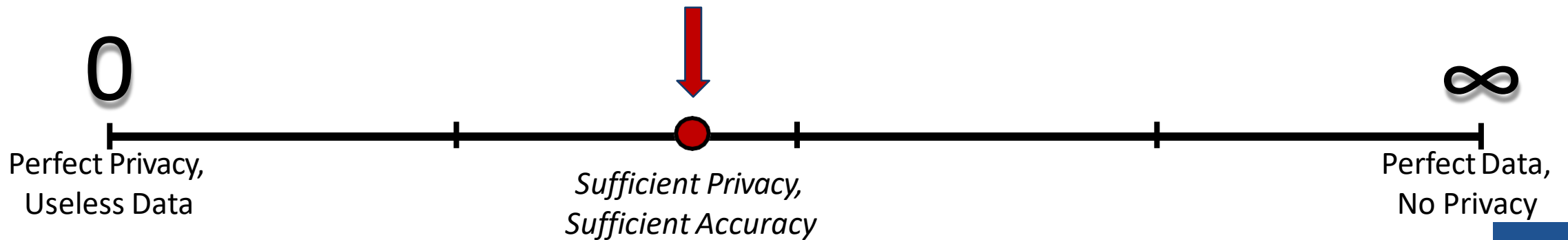
# What is a Privacy-loss Budget?

Determining the optimal PLB is a (difficult) policy decision



# What is a Privacy-loss Budget?

Comparisons to alternative methodologies can help put these trade-offs into perspective



# Background

**DAS Reconstruction Team efforts since February 2020**

Pre-decisional

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**

# Formation and goals of DAS Reconstruction group

- The DAS Science and DevOps team continue to finalize implementation of the TopDown Algorithm for 2020 Census production
- In February 2020, a group in CED-DA began assessing the potential impacts of swapping, using an algorithm based upon the one used for the 2010 Census
- This team has become the DAS Reconstruction team, and has since performed these swapping experiments and generated preliminary assessment of the impact of suppression



# Suppression

Experiments based upon 1980 Census suppression rules and OMB race categories

Pre-decisional

Shape  
your future  
START HERE >

United States<sup>®</sup>  
**Census**  
**2020**

# Suppression Primer

- Suppression involves removing information from published tables to protect privacy
- The 1980 Census used two types of suppression: table suppression and cell suppression
- Table suppression involves deleting tables that fail specified thresholds
- Cell suppression involves deleting individual table cells that fail specific thresholds
- Cell suppression is typically harder to implement due to the need for complementary suppression

# Suppression Primer: Complementary Cell Suppression

Variable A	Category 1	Category 2	
Variable B			
Category 1	20	17	37
Category 2	2	15	17
	22	32	54

Cell value is too small

Variable A	Category 1	Category 2	
Variable B			
Category 1	20	17	37
Category 2	S	15	17
	22	32	54

Suppress the value

# Suppression Primer: Complementary Cell Suppression

Variable A	Category 1	Category 2	
Variable B			
Category 1	20	17	37
Category 2	S	15	17
	22	32	54

Other cells and table margins allow  
recovery of suppressed value

Variable A	Category 1	Category 2	
Variable B			
Category 1	S	S	37
Category 2	S	S	17
	22	32	54

Complementary suppression prevents  
this from happening

# Suppression from the 1980 Census

- The DAS Reconstruction team assessed the impact of applying 1980 Census-based suppression rules to the P.L. 94-171 (redistricting data) and Summary File 1 products (the “Demographic and Housing Characteristics” (DHC) file in 2020) based on the 2010 Census Edited File (CEF)
- The team used race and ethnicity categories specified by the Office of Management and Budget in Statistical Policy Directive 15 (1997) and implemented by the Department of Justice Voting Section
  - White alone
  - Black alone or in combination with white
  - Asian alone or in combination with white
  - Native Hawaiian or other Pacific Islander alone or in combination with white
  - American Indian or Alaska Native alone or in combination with white
  - Some other race alone or in combination with white
  - Two or more races, except as explicitly noted in the categories above
  - Hispanic/Not-Hispanic

# Suppression from the 1980 Census

## P.L. 94-171 Redistricting Data

- Table Suppression: Whole tables were suppressed (not published) for geographies with between 1 and 14 persons in any of the race/ethnicity groups
  - Applied to two tables:
    - (P3) Race for the Population 18 Years and Over, and
    - (P4) Hispanic or Latino, and not Hispanic or Latino, by Race for the Population 18 Years and Over
- Cell Suppression: Cell counts of 1 or 2 were replaced by 0
  - Applied to two tables:
    - (P1) Race
    - (P2) Hispanic or Latino, and not Hispanic or Latino by Race

## Additional Summary File (SF1) Data

- Table Suppression: Whole tables that are not dedicated solely to race and ethnicity are suppressed if their geographies have between 1 and 14 persons.
- For all person-level tables

# Impact of Suppression Rules on Privacy Risk

- Suppression, if done correctly, removes information from the tables that are released
- This means that enough suppression done on a set of tables can prevent re-identification attacks based on reconstruction of microdata from those tables
- While this would eliminate the risk of a specific attack on a specific set of tables, it is not equivalent to the broad privacy protection associated with formal privacy definitions

# Suppression Results: P.L. 94-171

- Under the 1980 suppression rules, tables P1 and P2 would have cell suppression applied only
- Cells with counts of 1 or 2 would be reported as 0
- The population total margin of P1 and P2 is never suppressed
- *These results include only primary cell suppressions*
- *Complementary suppressions would be necessary to prevent recovering cell values from margins*

## P1: Race

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	7	0	0
State	357	0	0
County	22,001	530	2.4
Tract	507,717	28,024	5.5
Block Group	1,518,048	153,914	10.1
Block	43,449,189	3,538,888	8.1

DRB clearance number CBDRB-FY21-213

## P2: Hispanic or Latino, and Not Hispanic or Latino by Race

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	14	0	0
State	714	0	0
County	44,002	2,987	6.8
Tract	1,015,434	110,081	10.8
Block Group	3,036,096	440,539	14.5
Block	86,898,378	5,071,570	5.8

DRB clearance number CBDRB-FY21-213



# Suppression Results: P.L. 94-171

- Results of the experiment show that table suppression for P.L. 94-171 tables P3 and P4 would exceed 84% and 87% (respectively) for on-spine geographies below the county level (tract, block group, block)

## P3: Race For The Population 18 Years and Over

Geography	Total Tables	Suppressed Tables	% Tables Suppressed
Nation	1	0	0
State	51	0	0
County	3,143	1,610	51.2
Tract	72,531	61,177	84.3
Block Group	216,864	207,643	95.7
Block	6,206,505	5,204,047	83.8

DRB clearance number CBDRB-FY21-213

## P4: Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over

Geography	Total Tables	Suppressed Tables	% Tables Suppressed
Nation	1	0	0
State	51	0	0
County	3,143	2,645	84.2
Tract	72,531	72,346	99.7
Block Group	216,864	216,759	100.0
Block	6,206,505	5,445,153	87.7

DRB clearance number CBDRB-FY21-213

# Suppression Results: P.L. 94-171

- The team also assessed the potential impact of cell suppression on tables P3 and P4
- This would imply adding voting age as part of the cell suppression criteria
- *These results include only primary cell suppressions*
- *Complementary suppressions would also be necessary to prevent recovering cell values from margins*

## P3: Race For The Population 18 Years and Over

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	7	0	0
State	357	0	0
County	22,001	822	3.7
Tract	507,717	38,439	7.6
Block Group	1,518,048	204,853	13.5
Block	43,449,189	4,200,018	9.7

DRB clearance number CBDRB-FY21-213

## P4: Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over

Geography	Total Cells	Cells Changed to Zero	% Cells Changed
Nation	14	0	0
State	714	0	0
County	44,002	4,078	9.3
Tract	1,015,434	146,400	14.4
Block Group	3,036,096	533,314	17.6
Block	86,898,378	5,822,712	6.7

DRB clearance number CBDRB-FY21-213

# Suppression Results: SF1

- The team assessed the impact of table suppression on additional 2010 SF1 tables by counting how many geographies meet broad restrictions on the total population and housing units
- This assessment showed that suppression of SF1 at the block level would exceed 38% for person-level tables and 32% for housing unit tables
- Additional SF1 table suppressions would be necessary at the block group and tract levels as well

## SF1: Geographies meeting criteria for person table suppression

Geography	Total populated	Population meets criteria	% Meets Criteria
Nation	1	0	0
State	51	0	0
County	3,143	0	0
Tract	72,531	131	0.2
Block Group	216,864	204	0.1
Block	6,207,027	2,401,802	38.7

DRB clearance number CBDRB-FY21-213

## SF1: Geographies meeting criteria for housing table suppression

Geography	Total occupied	Housing unit count meets criteria	% Meets Criteria
Nation	1	0	0
State	51	0	0
County	3,143	0	0
Tract	72,425	182	0.3
Block Group	216,598	307	0.1
Block	6,188,078	2,027,988	32.8

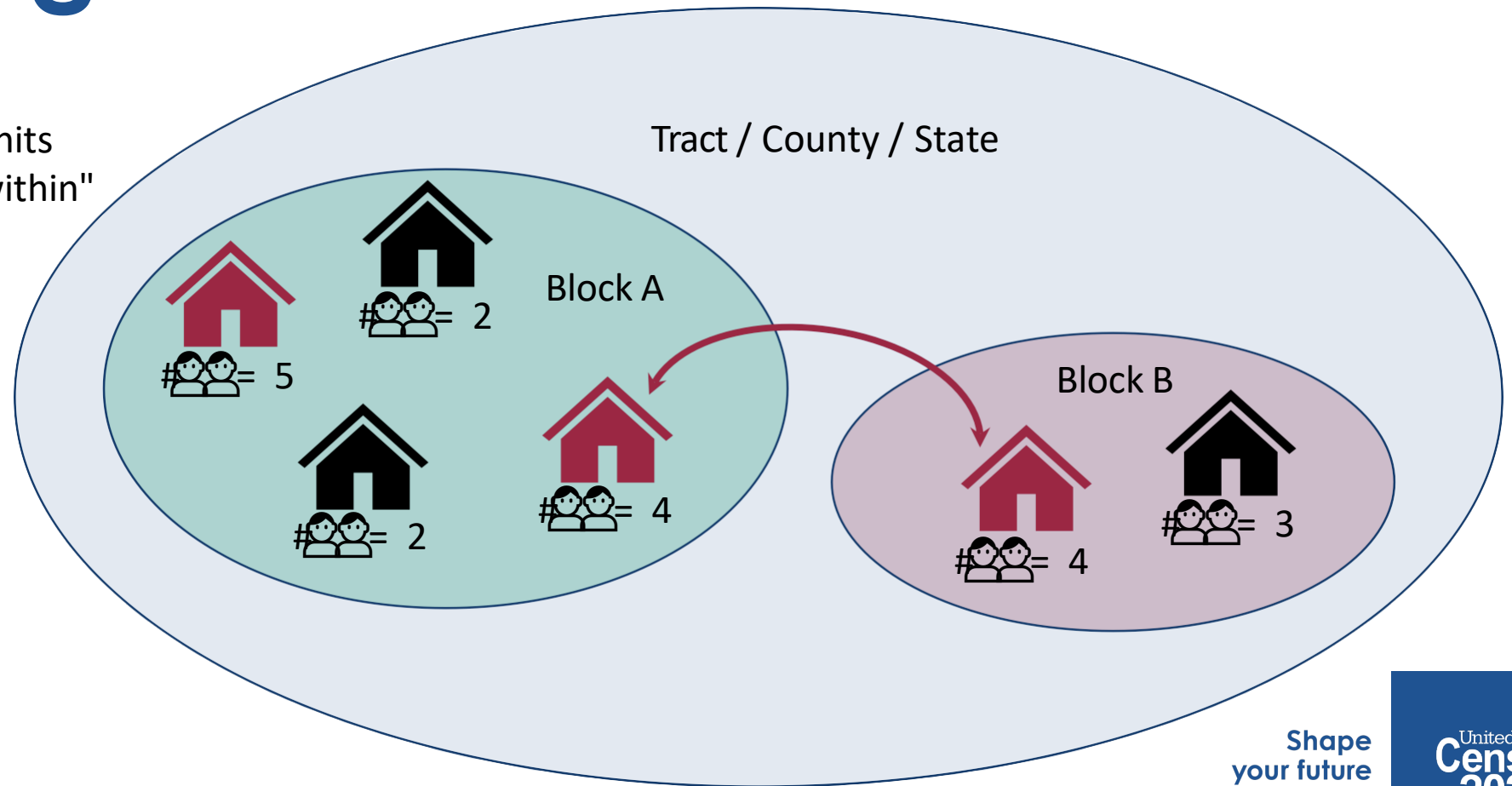
DRB clearance number CBDRB-FY21-213

# Swapping

Relaxations and extensions of the 2010 Census swapping algorithm

# Swapping Primer

1. Determine key to match units
2. Choose "between" and "within" geographies
3. Determine units to swap
4. Select swap rate
5. Find swap pairs



# Adapting the 2010 Swapping Algorithm for higher rates

- Initial efforts of the DAS Reconstruction team focused on adapting the 2010 Census swapping to support higher swap rates, up to 100% if necessary
- This algorithm now has the following parameters and adjustments:
  - The desired swap rate
  - The list of invariants (the swap “key”)
  - Mechanisms for relaxing invariants and extending swapping beyond tracts

# Swapping Experiments

- The DAS Reconstruction team has prepared swapped files for numerous iterations of the parameters
  - Swap rates ranging from 5% to 50% of housing units
  - Pre-swap perturbation of household size by  $\pm 1$  for up to 80% of housing units
  - Pre-swap perturbation of tract within county or within state for up to 70% of housing units
- At the beginning of CY2021, the team began to assess the impact of these parameters on the outcomes of the reconstruction-abetted re-identification attack on the 2010 Census

# Swapping Results

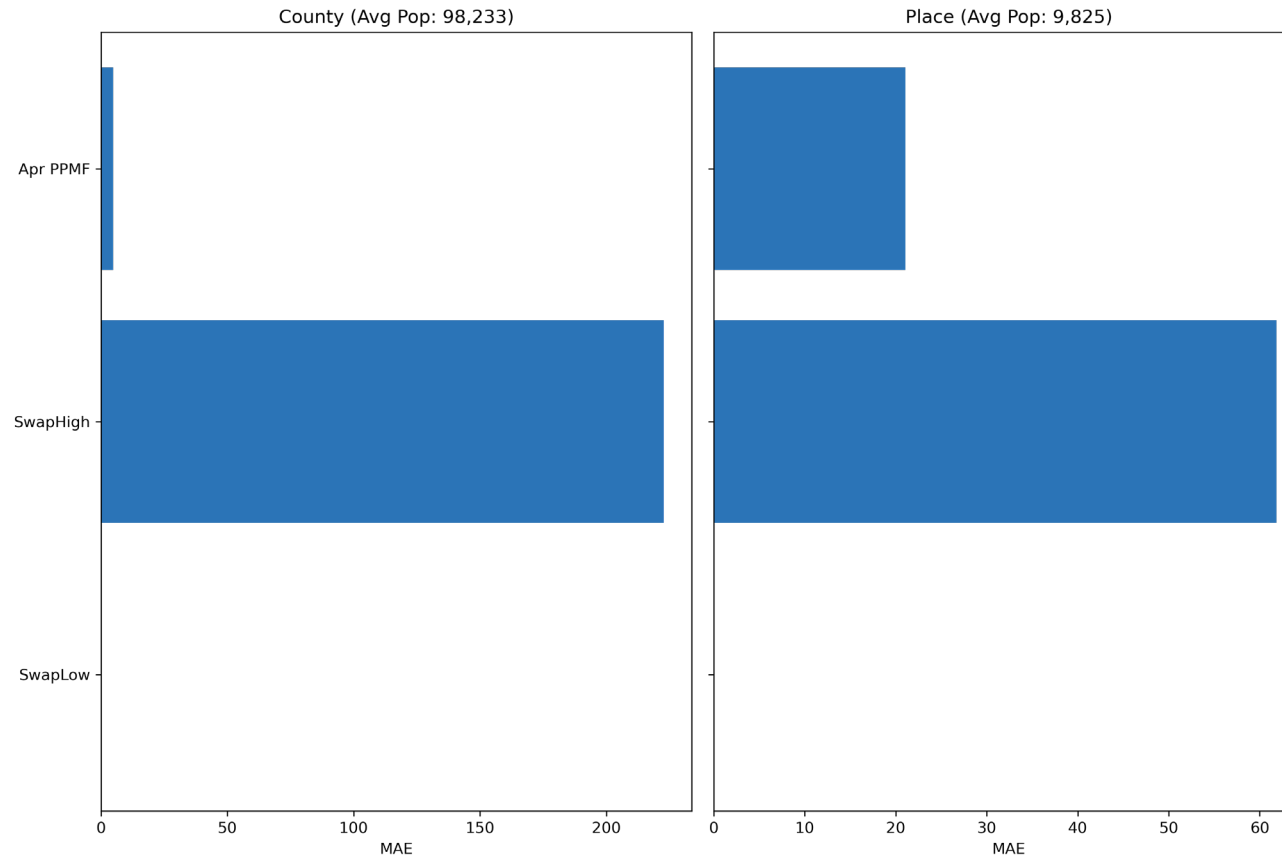
- The key swapping outcomes of those experiments have been:
  - Low swap rates have essentially no impact on re-identification outcomes; they are essentially the same as for the 2010 SF1
  - High swap rates have only a minimal impact on re-identification outcomes, with accuracy metrics inferior to the 4/28/2021 Disclosure Avoidance System (DAS) Privacy-Protected Microdata File (PPMF)
- These imply that middling swap rates, as implemented, may match the TopDown Algorithm in terms of accuracy but will have a low impact on reducing re-identification

Swap Parameters				Reidentification		
Experiment	Swap %	%HH Size Perturbed	%Tract perturbed	Putative % of Population	Confirmed % of Population	Precision (Confirmed/Putative)
2010 HDF	-	0	-	44.60	16.85	37.79
SwapLow	5	0	0	44.38	16.52	37.23
SwapHigh	50	50	70	42.69	12.96	30.37



# Swapping Results

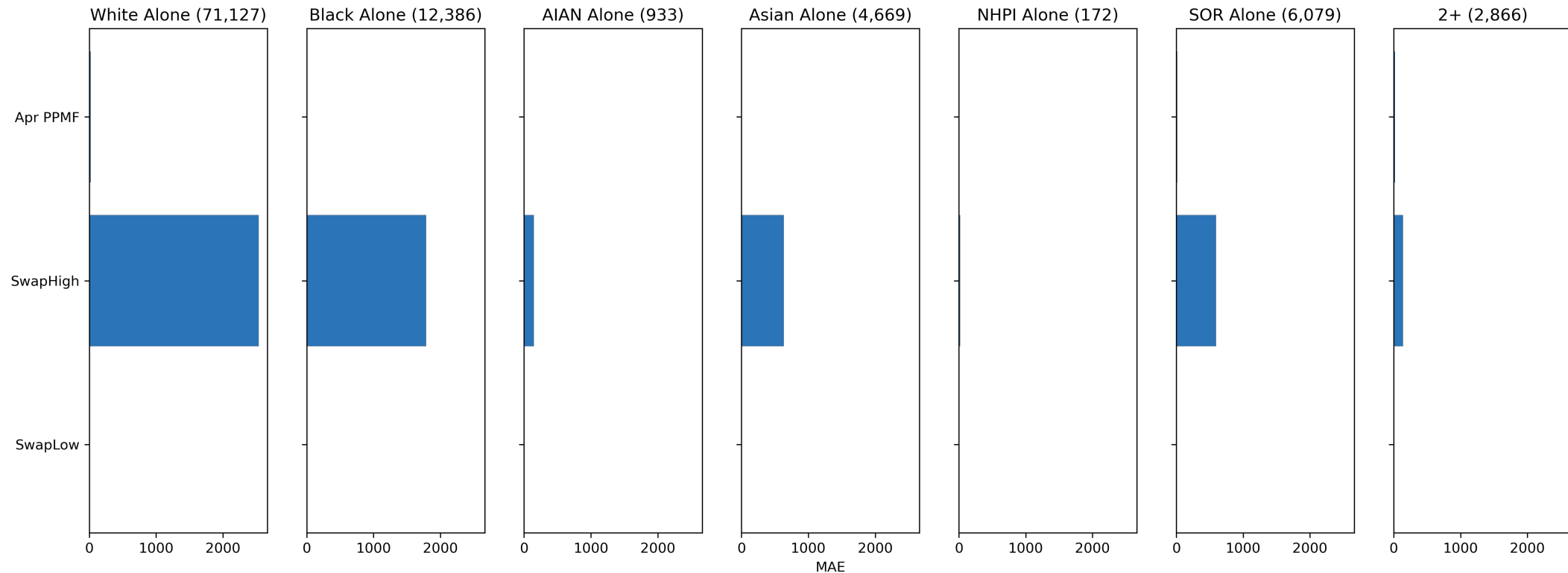
Comparison of mean absolute error (MAE) for total population for county and incorporated place size categories



DRB clearance number CBDRB-FY21-213

# Swapping Results

Comparison of mean absolute error (MAE) for race alone for counties



DRB clearance number CBDRB-FY21-213

# Final Considerations

- None of the algorithms described herein adheres to a formal definition or semantic for privacy loss, and they are only being assessed against one attack strategy (the 2010 Census reconstruction-abetted re-identification attack)
- Implementation of the 1980 Census suppression rules would lead to extreme amounts of table suppression for sub-state on-spine (county, tract, block group, block) geographies
- Implementation of relaxations and extensions of the 2010 Census swapping algorithm would yield little improvement in re-identification outcomes even at high swap rates
- Production implementation of either suppression or swapping is expected to take at least an additional 6 months after a decision to implement them